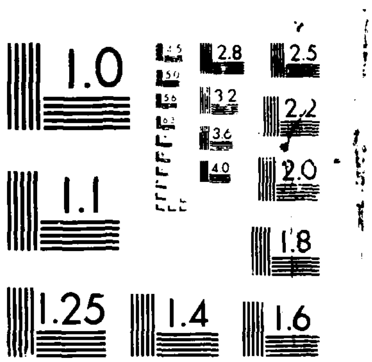
AD-A195 302 SMOOTHING SPATIAL DATA BY ESTIMATING MEAN LOCAL VARIANCE(U) NAVAL POSTGRADUATE SCHOOL MONTEREY CA L D JOHNSON APR 88 NPS55-88-005 1/1

UNCLASSIFIED F/G 24/1 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963

NPS55-88-005

NAVAL POSTGRADUATE SCHOOL

Monterey, California

SMOOTHING SPATIAL DATA BY ESTIMATING MEAN LOCAL VARIANCE

LAURA D. JOHNSON

APRIL 1988

Prepared for:
Naval Postgraduate School
Monterey, CA 93943-5000

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release; distribution is unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
NPS55-88-005

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Naval Postgraduate School

6b. OFFICE SYMBOL *(If applicable)*
Code 55

7a. NAME OF MONITORING ORGANIZATION
Naval Postgraduate School

6c. ADDRESS *(City, State, and ZIP Code)*
Monterey, CA 93943-5000

7b. ADDRESS *(City, State, and ZIP Code)*
Monterey, CA 93943-5000

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
Naval Postgraduate School

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
O&MN, Direct funding

8c. ADDRESS *(City, State, and ZIP Code)*
Monterey, CA 93943-5000

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.

11. TITLE *(Include Security Classification)*
SMOOTHING SPATIAL DATA BY ESTIMATING MEAN LOCAL VARIANCE

12. PERSONAL AUTHOR(S)
Johnson, Laura D.

13a. TYPE OF REPORT
Technical

13b. TIME COVERED
FROM ______ TO ______

14. DATE OF REPORT *(Year, Month, Day)*
1988 - April

15. PAGE COUNT
35

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
air pollutants, local variance, nearest-neighbor regression, smoothing, variogram

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

A nearest neighbor regression method is used to estimate air pollution levels at other than measured points. The method requires an appropriate smoother. Cross-validation is used to determine the appropriate smoother. An alternative method is introduced to determine an appropriate level of smoothing which involves minimizing mean local variance. Mean local variance is a function of the size of a circular window. It is minimized for two pollutants in Ohio, New York and Florida. The smoother obtained by cross-validation using Ohio's data is compared to that obtained by minimizing mean local variance.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
Laura D. Johnson

22b. TELEPHONE *(Include Area Code)*
(408)646-2569

22c. OFFICE SYMBOL
Code 55Jo

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

☆ U.S. Government Printing Office: 1986—606-243

1. Introduction

For simplicity, let us describe the problem in two dimensions. Assume that we have a physical variable $Z(x,y)$ in some two-dimensional region. Thus $Z(x,y)$ may be ozone concentrations in the San Francisco Bay Area region. Assume also that this random field Z(x,y) is measured at fixed points $x_i, y_i, i=1,2,...n$. The problem consists of estimating $Z(x,y)$ at arbitrary points in the region from the given data.

Nearest neighbor regression (NNR) functions are useful for estimating the value of $Z(x,y)$. Nadaraya (1964) and Watson (1964) first independently proposed a NNR method which behaves well even when only little information on the distribution of the measured data points is known. Stone (1980) showed that non-parametric regression estimators of this type have uniform rate of convergence. Stute (1984) showed that if the $E[Z^2]$ is finite, these estimates are asymptotically normal. The same year Cheng (1984) showed that if the noise about the true regression has finite variance these estimates are uniformly consistent. Stute (1984) also showed that these estimates are more efficient than kernel estimators when there are a limited number observations in the neighborhood of the estimated point.

Air Quality is monitored in the United States by monitoring apparatus of state, local, and federal networks. From 1974 to 1976, 5777 monitoring stations existed. These air quality data consist of discretely measured points which are not on a regular grid. Details regarding these data are given in (Johnson,1983). Section 2 introduces NNR as it is applied to these data. Cross-validation is used for choosing the smoothing parameter. Section 3 introduces a concept called local variance and its relationship to the necessary amount ot smoothing for a set of data. A measure of local variance as a func-

tion of window size is also introduced in this section. Section 5 and 6 finds the window size for which mean local variance is minimized for three states and two pollutants. The window size is then related to the smoothing parameter. The choice of smoothing parameter given by cross-validation is compared to that given by the minimum mean local variance method.

2. The Nearest Neighbor Regression Method

Let Z_i be the information from data points positioned at (x_i, y_i). For example, the value of Z_i is the measured pollutant level at latitude x_i and longitude y_i. Let (x_p, y_p) be the point of interest of a neighborhood P. Z_i's may or may not be located in P. Define d_i as the euclidean distance of (x_i, y_i) to (x_p, y_p). If d_i is large we want Z_i to be less influential in the estimate of Z_p than if it were small. Let $m(x_p, y_p) = E(Z_p / x_p, y_p, Z_1, Z_2, \cdots Z_n)$. We are required to construct an estimate of $m(x_p, y_p)$. The proposed estimate is

$$m^*(x_p, y_p) = \sum_{i=1}^{n} \lambda(d_i) Z_i$$

where $\lambda(d_i)$ is based on an appropriate weight function and the $\lambda(d_i)$ sum to one. More specifically,

$$\lambda(d_i) = \frac{\omega(d_i)}{\sum_{i=1}^{n} \omega(d_i)}$$

where $\omega(d_i) = e^{-0.5d_i^2/d_o^2}$. Since some stations are more active than others: if p_i is the percentage of time the station located at (x_i, y_i) was active, then we use $\omega(d_i) = p_i e^{-0.5d_i^2/d_o^2}$.

The weights are dependent on distance such that they are relatively large for measurements within a relatively small radius of the estimated point. The size of d_o determines how relatively large the weights are. Weights are large for large d_o and small for small d_o. As d_o increases the amount of smoothing increases. Figure 1. gives shows the shape of the weight functions for constant p_i. Each curve corresponds to a different d_o which covers 1, 2, 5, 10, 20, and 50 km. Note that the flattest curve in Figure 1. is for d_o=50*km* and thus gives the greatest amount of smoothing; while the steepest curve is for d_o=1*km* and gives the least amount of smoothing.

These weights were used in the Populations at Risk to Environmental Pollution (PAREP) project for estimating pollution concentration at the population centroid in various geographic areas. According to Merrill (1982), these weights were chosen for three reasons "(i) the estimated function would be smooth in the vicinity of the estimated points; (ii) the estimated function need not pass through all measured points; (iii) the area integral of the estimated function should be finite, so that distant points can be ignored in the calculation."

To avoid producing estimates from poorly monitored stations, some constraints in estimation are used. The constraints are as follows. Let $(x_{(1)}, y_{(1)})$ denote the nearest data point to (x_p, y_p). It follows that $d_{(1)}$ is the minimum distance of a data point to (x_p, y_p).

If $d_{(1)} \geq 3d_o$ exclude $(x_{(1)}, y_{(1)})$ and conclude that Z_p is inestimable.

If $d_{(1)} < 3d_o$ then all Z_i within $4d_o$ will be used to estimate Z_p.

The $4d_o$ criterion has been constructed to avoid sharp discontinuities at the midpoint between predicted points which are exactly $6d_o$ apart. This is especially important for estimating a surface for which this method may also be used.

One drawback to this method is that these estimates are easily biased to clusters of data points. If the point of estimation is equidistant from 10 clustered points on one side and 1 point on the other side, the estimate using this procedure will be dominated by the 10 clustered points when the information from the lone point on one side may be more valuable since it is the only measured point on that side. A weight which is inversely proportional to the number of points within a specific distance of the observed point could be added to account for the clustering. Thus, points in clusters have less influence than those lone points on the estimate. Regardless of whether the weighting is by distance or cluster presence a smoothing parameter to determine the amount of smoothing would be included and thus methods for choosing the appropriate smoother are still applicable.

2.2 Cross-Validation to Determine the Amount of Smoothing

To choose the "optimum" value of d_o, a pollution estimate for each station location was produced from observations at other nearby stations excluding the station's own observed value. This estimate is then compared to the actual observed value at the station location of the estimate for various d_o's. By varying d_o, we can choose those estimates which minimize the prediction error. This method is called cross-validation and is nearly identical to the application in Wahba and Wold (1975). However, they used

only one measure of prediction error and in this paper four measures are considered. The idea to leave out one data point out at a time and see how well the various estimates fit the observed values is described fully in Stone (1974) or Geisser (1975).

Prediction error is defined as the cross-validatory estimate minus the observed value. Wahba and Wold (1975) use the mean of the squared prediction errors and call this the cross-validation mean square error technique. This technique is used to estimate from the data the appropriate degree of smoothing. In this paper, several composite measures of prediction error are used.

Since any composite function of the prediction errors is heavily biased by outliers, the technique should be used vigilantly. Although, the possibility of this bias is not explored here, many points are estimated and compared so that the influence of 1 or 2 outliers should not be substantial.

The composite prediction errors were calculated as follows. Let

z_i = logarithm of the pollutant concentration at location (x_i, y_i).

m_i^* = estimate of $Z(x_i, y_i)$ from surrounding data points.

n = the number of stations that could be estimated.

p_i = the percentage of time stations i was active.

The concentrations are in geometric means which are assumed to have a lognormal distribution. Thus, the logarithm of these values is distributed normally for large n. Therefore when observed or estimated values are discussed they are in units of the logarithm of the geometric mean concentration.

The reliability of this estimate is to some extent a function of the station activity as well as the estimation procedure. For these error functions to reflect the adequa-

cy of the estimation procedure and not the station activity they are weighted according to the percentage of time active.

The sample mean value of the pollutant level, Z is

$$\bar{z} = \sum_{i=1}^{n} z_i .$$

The sample variance of all the values of pollutant level, Z is

$$S_z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2.$$

The weighted variance of these values is

$$S_{zw}^2 = \frac{\sum_{i=1}^{n} p_i (z_i-\bar{z})^2}{\sum_{i=1}^{n} p_i}.$$

The following error measures have been considered.

(MSE) Mean Squared Error

$$MSE = \frac{\sum_{i=1}^{n}(z_i-m_i^*)^2}{n};$$

(WMSE) Weighted Mean Squared Error

$$WMSE = \frac{\sum_{i=1}^{n}p_i (z_i-m_i^*)^2}{\sum_{i=1}^{n}p_i};$$

(RMSE) Ratio of Mean Squared Errors

$$RMSE = \frac{1}{nS_z^2} \sum_{i=1}^{n}(z_i-m_i^*)^2;$$

(WRMSE) Weighted Ratio of Mean Squared Errors

$$WRMSE = \frac{1}{\sum_{i=1}^{n} p_i \ S_{zw}^2} \sum_{i=1}^{n} p_i (z_i - m_i^*)^2;$$

(MAE) Mean Absolute Error, Percent

$$MAE = \frac{1}{n\bar{z}} \sum_{i=1}^{n} | z_i - m_i^* | \times 100;$$

(WMAE) Weighted Mean Absolute Error, Percent

$$WMAE = \frac{1}{\bar{z} \sum_{i=1}^{n} p_i} \sum_{i=1}^{n} p_i | z_i - m_i^* | \times 100;$$

(MPE) Mean Percent Error

$$MPE = \sum_{i=1}^{n} \frac{p_i | z_i - m_i^* |}{n z_i} \times 100$$

(WMPE) Weighted Mean Percent Error

$$WMPE = \frac{\sum_{i=1}^{n} \frac{p_i | z_i - m_i^* |}{z_i}}{\sum_{i=1}^{n} p_i} \times 100$$

These formulas are similar to those used by Breiman (1977) for comparing kernel and Parzen multivariate density estimation techniques. Wahba and Wold (1975) use mean squared error in their presentation.

2.3 Smoothing Sulfur Dioxide and Suspended Particulate Data for Ohio State By Cross-Validation

Sulfur dioxide and suspended particulate data for the state of Ohio were used to illustrate the selection of an optimal d_o by cross-validation. From 1974 through 1976 there were 185 active stations for sulfur dioxide and 398 active stations for total suspended particulate. For d_o = 1, 2, 5, 10, 20 and 50 km, the NNR method was used to predict the observed value at the location of each of the active stations by leaving out the station's own value. The difference between the prediction and the observed value is the prediction error. We would like to choose the d_o with the smallest composite prediction error over all observed values.

The eight composite measures of error (explained in section 2.2) were computed for total suspended particulate and sulfur dioxide data in Ohio. Table 1. shows the value of each function for suspended particulate and different values of d_o. The value of d_o which minimizes the loss for the the majority of the unweighted error functions is 5 kilometers. The RSE is the only one which is not consistent with this choice of an optimal d_o. Whereas, the weighted error functions consistently choose d_o equal to 2 km.

TABLE 1. Error Functions by d_o (km) for Suspended Particulate in Ohio State.

d_o (km)	MSE	RSE	MAE	MPE	WMSE	WRSE	WMAE	WMPE
1	.07562	*.500	4.9	4.9	.0633	.425	4.5	4.4
2	.06907	.539	4.8	4.9	*.0519	*.404	*4.2	*4.1
5	*.06586	.621	*4.8	*4.0	.0531	.477	4.3	4.2
10	.06945	.696	4.9	4.9	.0565	.533	4.4	4.3
20	.08111	.836	5.3	7.4	.0673	.631	4.7	4.6
50	.09337	.963	5.6	5.7	.0836	.803	5.7	5.2

*indicates the minimum value and thus corresponds to the optimum d_o.

Table 2. is the analogue to Table 1. for sulfur dioxide data. Note that the both weighted and unweighted RSE is greater than one fo all d_o. This means that the sum of squared error of the estimate derived from the NNR method is greater than the sum of squared error using the overall mean. In other words,

$$\sum_{i=1}^{n}(z_i - m_i^*)^2 > \sum_{i=1}^{n}(z_i - \bar{z})^2.$$

Since the estimate contains only the 'nearest' points and the mean contains all the points in the region, this could arise from measuring problems such as a large measurement error or biased measurements. It could also be that the NNR method is inherently wrong for estimating sulfur dioxide. Unlike suspended particulate, sulfur dioxide is a specific chemical compound which may diffuse or transform to other compounds in the atmosphere. Therefore SO_2 concentrations may vary more in a smaller neighborhood (i.e., have higher local variance). Suspended Particulate includes many compounds of a certain size (eg. dust) which may not be as apt to change in concentration locally.

TABLE 2. Error Functions by d_o for Sulfur Dioxide in Ohio State.

d_o (km)	MSE	RSE	MAE	MPE	WMSE	WRSE	WMAE	WMPE
1	4.34	4.42	57.0	54.9	3.49	3.91	46.6	43.1
2	2.52	*3.56	39.9	38.2	1.69	*2.42	*27.7	*25.8
5	2.17	3.98	38.7	38.1	1.63	2.82	28.9	27.1
10	2.11	4.21	39.8	39.9	1.60	2.93	30.6	29.5
20	2.00	4.32	39.7	39.8	1.56	2.96	31.1	29.8
50	*1.80	3.95	*38.2	*37.5	*1.38	2.68	*37.5	29.7

* indicates the minimum value and thus corresponds to the optimum d_o

3.1 Determining the Appropriate Smoother by Estimating Local Variance

The appropriate amount of smoothing should be such that insignificant variance among adjacent stations is smoothed out; and yet significant trends over larger geographic areas are preserved. Thus, perhaps a more direct method than cross-validation for finding the optimal d_o or a range of optimal d_o's is to construct a measure of the mean local variance (MLV) as a function of d_o for a specific area and find the d_o which minimizes it. That is, in the NNR method the d_o which is chosen should be approximately that d_o, corresponding to a distance from each point, for which the point values within this distance are, on the average, the most homogeneous. This section explores a method for measuring mean local variance as a function of distance. The relationship between d_o and distance is explained below.

Since the weights are proportional to

$$\omega(d_i) = e^{-.5 d_i^2 / d_o^2},$$

points which are further away relative to d_o get less weight than points close to the estimated point. Therefore, for a particular spatial distribution of points, d_o can be

thought of as a distance at which some multiple of it makes the weights so small that the station's information receives insignificant weight. Furthermore, only points within four times d_o are used in the average; yet the points that are a distance of four times d_o give an absolute weight of

$$\omega(4d_o) = e^{-8} = .0003$$

which can be a small or large relative weight. Since the weights in this method are relative weights, a point which is far away, say 2.9 d_o, can have a weight=1.0 if it is the only point within $3 \times d_o$. Therefore choosing a relationship between distance and d_o for this weighting scheme depends on the particular data set and its precision. 't is suggested here to estimate the mean local variance for a particular data set as a function of window size and choose the appropriate window size on this basis. The relationship between window size and d_o is positive but will vary among different data sets.

Local variance is defined as the variance in the data within a small neighborhood. Local variance is measured in the following manner. A circular window is drawn around every point, defined by the station location, in the entire geographic area of interest. Thus, for every point there is an associated circular window for which it is the center. The radius of each window is chosen as a function of d_o and is allowed to vary from 0 to encircle the entire area. For a large enough radius there will be other points besides the center point contained inside these circular windows. For each point we can calculate a mean and a variance using the points which lie in its associated circular window for a specified radius r. For example, denoting k_i as the number of points within z_i's associated window and z_j as one of those points contained in the window where j ranges from 1 to k_i, the window associated with z_i has the following local mean and variance.

Local Mean:

$$\bar{z}_i = \frac{\sum_{j=1}^{k_i} z_j}{k_i};$$

Local Variance:

$$\sum_{j=1}^{k_i} \frac{(z_j - \bar{z}_i)^2}{k_i - 1}.$$

To calculate the local variance of the entire region for a particular r, an average of the local variances is taken and is called the mean local variance. Local variance as a function of distance is obtained by varying the size of the circular window.

First, let us decide how to weight the local variances to find the mean local variance for a particular data set. Since a point may be in more than one window we should weight the local variances so that each point contributes a weight of 1 to the average local variance for the entire region. Each window can be thought of as having $k_i - 1$ "degrees of freedom" where k_i is the number of points in each circular window. If each window were independent of every other window (i.e. non-overlapping groups of points) then the weights would be 1 and mean local variance could be measured by the unweighted average of the local variances. But since there is apt to be overlap, each squared difference associated with each point is given a weight of $\frac{1}{n_i}$ where n_i = the number of circular windows containing point (x_i, y_i). With this weighting, each point contributes a weight of 1 to the mean local variance of the region.

In order to calculate mean local variance we need a denominator which expresses the number of independent data points in the sum of local variances. We will

call this the "degrees of freedom". The "degrees of freedom" should take account of both the total number of data points and the amount of overlap in the circular windows. In fact, if each point contributes $\frac{1}{n_i}$, then the number of non-independent terms in the numerator is $\sum_{i=1}^{N} \frac{1}{n_i}$. Therefore we have chosen the denominator to be $N - \sum_{i=1}^{N} \frac{1}{n_i}$ because $\sum_{i=1}^{N} \frac{1}{n_i}$ is the number of non-independent terms in the numerator. This gives the expected answer in a few simple cases which are given below.

From the above discussion, we have constructed the following measure.

Mean Local Variance (MLV)

$$MLV = \frac{\sum_{j=1}^{N} \sum_{i=1}^{k_j} \frac{(z_i - \bar{z}_j)^2}{n_i}}{N - \sum_{i=1}^{N} \frac{1}{n_i}}$$

where

N = the total number of points in the region of interest.

z_i = the value of pollution at the point (x_i, y_i).

$\bar{z}_j$ = the mean value corresponding to the circular window associated with z_j.

k_j = the number of points enclosed by z_j's circular window of which z_i is one of them.

Let us examine this formula in light of a few examples.

CASE 1. Every point is in only one circular window. In this case $n_i = 1$ and thus the denominator of MLV is zero. Since the radius is so small, no local variances can be measured. MLV is undefined in this case.

CASE 2. The radius is large enough that every point lies in every other point's circular window. Then n_i=N and k_j=N and

$$MLV = \frac{\sum_{j=1}^{N}\sum_{i=1}^{N}\frac{(z_i-\overline{z}_j)^2}{N}}{N-\sum_{j=1}^{N}\frac{1}{N}} = \frac{\sum_{j=1}^{N}(z_j-\overline{z})^2}{N-1}$$

which is simply the sample variance of the points for the entire region. In other words, MLV approaches the global sampling variance as the radius increases.

CASE 3. Suppose there are n_i completely overlapping circular windows. In other words, there are m discrete clusters of points. In this case each cluster has its own mean and sum of squares associated with it. Therefore each point has within a particular cluster the same mean, $\overline{z}_j$ and sum of squares associated with it as any other point in its cluster. In this case $n_i = k_j$, where k_j is the number of points in the jth cluster, and

$$MLV = \frac{\sum_{j=1}^{N}\sum_{i=1}^{k_i}\frac{(z_i-\overline{z}_j)^2}{k_j}}{N-\sum_{i=1}^{N}\frac{1}{k_j}} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{k_i}(z_i-\overline{z}_j)^2}{N-m}.$$

This is similar to the within mean square in one way analysis of variance (Scheffe', 1959).

CASE 3a. As a specific case of m discrete clusters, suppose each point has only one other point in its circular window. So there are a series of two overlapping windows which do not contain any other points. In this case n_i = 2 for all i and

$$MLV = \frac{\sum_{j=1}^{N}\sum_{i=1}^{2}\frac{1}{2}(z_i-\overline{z}_j)^2}{N-\frac{N}{2}} = \frac{\sum_{j=1}^{N/2}\sum_{i=1}^{2}(z_i-\overline{z}_j)^2}{N-\frac{N}{2}}.$$

This is case 3 with two points in each cluster. This generalises to the case where every point is in k windows which do not overlap, then

$$MLV = \frac{\sum_{j=1}^{N/k}\sum_{i=1}^{k}(z_i - \bar{z}_j)^2}{N - \frac{N}{k}}.$$

Since every circular window's centered point is in every other point's circle in the above examples of geographically clustered points, $k_j = n_i$. For any particular $(z_i - \bar{z}_j)^2$, z_i may be contained in n_i different circular windows while z_j may be contained in k_j and thus n_i and k_j are not necessarily equal. In other words, there may be windows that z_i is contained in but all the points in those windows are not necessarily contained in z_i's window.

Mean Local Variance is similar conceptually to the variogram function described in Journel and Huigbregts (1978). The variogram describes variability between two quantities separated by a distance h; while mean local variance describes the variability in the data within a radius h. Like the variogram, to estimate mean local variance from the realizations of $Z(x,y)$, it must be assumed that the variogram function depends only on h and not on the location of $Z(x,y)$. This implies second-order stationarity which is described in detail in Journel and Huigbregts (1978). Although the variogram function may also be useful for finding the appropriate smoother or window size in NNR, this is not explored in this paper. However, we will explore the use of mean local variance for this purpose and the method using the variogram is analogous.

Mean local variance is relevant to the choice of the smoothing parameter in the NNR method described earlier. If d_o is to be large, then the minimum mean local variance should occur at a larger radius so that relatively large weight is given to points relatively far from the point of interest. In other words, the homogeneity of points

encompasses, on the average, a larger area for larger d_o. Likewise, if d_o is chosen to be small, points far away will be given relatively small weight and therefore the minimum mean local variance should occur at a smaller circle radius.

Ohio is chosen as an example so that we can compare our minimum mean local variance choice of d_o with our cross-validatory choice of d_o in the previous section. Mean local variance is calculated for various radii. For each d_o, two associated radii were chosen. One was $4 \times d_o$ which is the same size that the NNR method uses in selecting stations to compute its estimate of a point. The other radius is chosen such that the $\omega(d_i)$ are equal to .01 at the edge of the circular window. In Figures 2. and 3. the log (base e) mean local variance is plotted against the radius of the circular windows for Suspended Particulate and Sulfur Dioxide respectively. Logarithms are taken to magnify the lower values of the curve. In Table 3. the actual values calculated for Ohio are given. For comparison with the cross-validatory choices of the previous section, we chose windows corresponding to d_o = 1, 2, 5, 10, 20, and 50 km for calculating MLV. The radius which gives the most homogeneous window size corresponds to a d_o = 2 km by this method. These results are the same as those given by cross-validation using weighted loss functions and similar to that obtained using the unweighted loss functions (Table 1.).

TABLE 3. Mean Local Variance (MLV) as a function of d_o for Suspended Particulate and Sulfur Dioxide in the State of Ohio.

d_o (km)	radius (km)	Suspended Particulate		Sulfur Dioxide	
		MLV	$N-\sum_{i=1}^{n}\frac{1}{n_i}$	MLV	$N-\sum_{i=1}^{n}\frac{1}{n_i}$
1	3.03	.0583	141.80	.2129	50.37
1	4.00	.0573	173.28	.1934	66.70
2	4.29	*.0562	181.37	.1984	71.00
2	8.00	.0597	240.35	.1930	101.46
5	10.74	.0614	264.68	*.1887	113.73
5	20.00	.0662	326.96	.1989	135.56
10	21.47	.0671	334.21	.2008	137.58
10	40.00	.0819	372.48	.2526	162.65
20	42.94	.0837	375.20	.2644	164.31
20	80.00	.0927	390.68	.3166	177.30
50	107.36	.0945	393.21	.3468	180.41
50	200.00	.0957	396.03	.4063	183.10
**	500.00	.0972	397.00	.5069	184.00

* The smallest mean local variance, corresponding to the optimal d_o.
** d_o is that value which gives a radius large enough to contain all the points in Ohio. Notice that "df"=N-1 since there are 398 active stations for suspended particulate and 185 active stations for sulfur dioxide in the region.

In section 2.3, the cross-validatory analysis for sulfur dioxide using weighted loss functions suggests a d_o = 2 km while the unweighted analysis leans toward a d_o = 50 km. This implies that sulfur dioxide is more locally heterogeneous than suspended particulate. Minimum mean local variance analysis shows this is true. Table 3. gives a minimum at a circle radius of 10.74 km which corresponds to d_o = 5 km. Notice that this solution creates no conflict of choice between 2 d_o's which are so far apart. The minimum radius chosen will be at least near the global minimum.

In both figures 2. and 3., MLV increases after a minimum area and continues to increase steadily to the global sample variance of the entire region. Note the slight instability of the curve at smaller circular window size. This instability is expected since

with small k_i in each window we expect MLV to be extremely sensitive to the addition of new points with slight increments in window size.

3.2 The Shape of Mean Local Variance

Mean Local Variance was calculated for sulfur dioxide and suspended particulate in the states of New York and Florida. The results are revealed in this section. New York and Florida were chosen because, out of all states in the United States, these states have the second and third largest number of monitoring stations respectively. For each state, the basic shape of the curve showing mean local variance against circular window size is similar. Sulfur dioxide is consistently more locally heterogeneous than suspended particulate across all three states. (Tables 3., 4. and 5., along with Figures 2. thru 7.)

TABLE 4. Mean Local Variance (MLV) as a function of d_o for Suspended Particulate and Sulfur Dioxide in New York State.

d_o (km)	radius (km)	Suspended Particulate		Sulfur Dioxide	
		MLV	$N-\sum_{i=1}^{n}\frac{1}{n_i}$	MLV	$N-\sum_{i=1}^{n}\frac{1}{n_i}$
1	3.03	.0484	75.69	.2023	40.18
1	4.00	.0480	92.47	.1851	48.51
2	4.29	.0493	98.27	.1845	51.06
2	8.00	*.0481	132.76	.1615	71.95
5	10.74	.0516	154.79	*.1614	83.07
5	20.00	.0608	202.26	.2080	103.33
10	21.47	.0631	208.31	.2079	104.21
10	40.00	.0808	245.42	.2588	114.86
20	42.94	.0837	248.50	.2608	115.95
20	80.00	.0974	265.88	.3085	124.62
50	107.36	.1033	269.99	.3537	128.44
50	200.00	.1107	273.30	.3739	131.49
**	700.00	.1168	275.00	.3839	133.00

TABLE 5. Mean Local Variance as a function of d_o for Suspended Particulate and Sulfur Dioxide in Florida State.

d_o	radius (km)	Suspended Particulate		Sulfur Dioxide	
		MLV	$N-\sum_{i=1}^{n}\frac{1}{n_i}$	MLV	$N-\sum_{i=1}^{n}\frac{1}{n_i}$
1	3.03	.0580	48.73	.2365	46.82
1	4.00	*.0546	69.20	.2308	59.74
2	4.29	.0563	72.56	.2298	62.00
2	8.00	.0574	105.35	*.2176	81.64
5	10.74	.0608	122.99	.2325	90.68
5	20.00	.0654	151.83	.2426	105.83
10	21.47	.0668	155.72	.2423	107.50
10	40.00	.0716	176.02	.3138	118.06
20	42.94	.0716	177.97	.3151	118.75
20	80.00	.0841	191.81	.3222	127.46
50	107.36	.0866	194.26	.3452	131.39
50	200.00	.0918	200.73	.3595	134.17
**	500.00	.0907	201.00	.4017	184.00

* The smallest mean local variance, corresponding to the optimal d_o.
** d_o is that value which gives a radius large enough to contain all the points in the region.

States and pollutants have different window sizes that give minimum mean local variance. Since there are different dispersion and dilution mechanisms in different geographic areas, in addition to the different spatial distribution of the samples we should expect different choices across states and pollutants. The table below shows the radii at which the minima occur among states and pollutants.

State	Radius (km) Suspended Particulate	Radius (km) Sulfur Dioxide
Ohio	4.29	10.74
New York	8.00	10.74
Florida	4.00	8.00

In Figures 4. through 7., the line graphs are drawn to see the shape of MLV and there are basically three regions on each curve. The first region is called the Undefined Region. This area ranges from radius of 0 to a radius large enough to have at

least one circular window containing two points. The size of the undefined region is determined by the particular geographic area one is studying and the spatial distribution of stations in that area. The second region is called the Region of No-Precision. The Undefined Region is included in the Region of No-Precision. This region corresponds to circular windows which are too small to have enough points to measure local variance. This area is not specifically defined but can be estimated by looking at the "degrees of freedom" and plots of the logarithm of the radius vs. MLV. The "degrees of freedom" specify the amount of overlap and therefore how many points lie in each other's corresponding windows. The Region of No-Precision is designated in each of the plots, yet the boundary of this region is not sharp. The third region is for radii greater than those in the no-precision region. Window sizes in this region are large enough to contain enough points so that local variance can be measured. It is in this area that the optimum choice for d_o exists.

For the sake of example, plots of the "degrees of freedom (df)" for Ohio state's suspended particulate against logarithm of the window radius are given in Figure 8. Note that "df" increases quickly when radii are small and then levels out as they get large enough to include almost every station in the area. Degrees of freedom are a monotonic function of window size. The plot of "df" versus the log of the radius is an S-shaped curve. It is within the steep part of the S that the minimum MLV is likely to occur.

It is relatively easy to choose the optimum circle radius if MLV has the functional behavior shown in most of these curves. In Figure 4. it is more difficult because MLV does not increase steadily but bounces around initially. Using the method described in the previous paragraph, we would choose the minimum, which in this case,

lies in the area of no-precision, since it is the first trough-like area after a rapid increase. This window radius is only one kilometer. Knowing something about the spatial distribution of monitoring stations, it is seen that one kilometer would not allow many stations to have enough points contained in their windows to measure local variance. With this choice, the area of no-precision would have to end at a radius of .5 kilometers corresponding to an MLV with 8 "degrees of freedom". This implies that $\sum_{i=1}^{N} 1/n_i = 268$. Since there are only 276 stations altogether, n_i must be 1 for most stations. Thus at a circle radius of .5 kilometers, there are an insufficient number of points to measure mean local variance and thus it should be considered imprecise at this point. Therefore, care must be taken in the choice of local minimum. We must closely examine the degrees of freedom to see how reliable MLV is at various circular window sizes before choosing the minimum.

Perhaps there are two ways one can use mean local variance to choose the "best" circular window radius corresponding to the "best" d_0. One method is to choose a point estimate, the minimum radius greater than the "no-precision" area's upper limit. This method might not be the best since the MLV curve is data dependent, and therefore fluctuates around the minimum making it difficult to determine. Another involves choosing a region of radii called the "Region of Homogeneity". In each of the curves of MLV, after the region of no-precision there is a trough-like area in which the minimum occurs and then MLV starts to increase rapidly. We could call this trough-like area the Region of Homogeneity. Choosing a larger radius allows us to use more points for reliability. Yet, we do not want to use insignificant points. Thus choosing the largest possible radius before MLV starts increasing rapidly towards the global sample variance may be another acceptable method. In other words, one should not take any points into the estimate

which are outside the area of homogeneity. Which point is chosen as giving the most homogeneous window size is not critical for estimation. However, it is important to know which window sizes not to use, and those sizes can be determined easily by investigating MLV and its relationship with circular window size.

4. Conclusions

A non-parametric regression method has been used to estimate pollution concentrations for particular geographic areas. To determine the best value of the smoothing parameter d_o, cross-validation techniques were used. Cross-validation has been used as a method for finding the appropriate level of smoothing by the work of Wahba (1975) and others. Use of the concept of mean local variance (MLV) is explored as an alternative means of finding d_o. An example of the MLV approach has been given here using these data. Further work would include simulations of $Z(x,y)$ for finding the distributional properties of mean local variance.

Minimum mean local variance analysis could also be used for choosing parameters in other methods of spatial estimation. For example, mean local variance can be plotted as a function of the number of nearest neighbors or the size of the partition and thus the minimum mean local variance could correspond to the optimum number of neighbors or partition size in the same way it does d_o.

Mean Local Variance is also of interest in its own right. MLV can be useful for comparing geographic regions, spatial sampling distributions and different pollutant's dispersion mechanisms.

References

1. Bartlett, M. S. (1975) *The Statistical Analysis of Spatial Pattern.* Chapman and Hall, London.

2. Breiman, Leo, Meisel, W. and Purcell, E. (1977) Variable Kernel Estimates of Multivariate Densities. *Technometrics, Vol. 19, no. 2.*

3. Cheng, P.E. (1984) Strong consistency of nearest neighbor regression function estimates. *J. Mult. Analysis, 15, 63-72.*

4. Diggle, P.J. (1983) *Statistical Analysis of Spatial Point Patterns.* New York: Academic Press.

5. Efron, Bradley (1982) The Jackknife, the Bootstrap and Other Resampling Plans. *CBMS-NSF Regional Conference Series in Applied Mathematics,* 38, 49-59.

6. Geisser, Seymour (1975) The predictive sample reuse method with applications. *J. American Stat. Assoc., 70, 320-328.*

7. Johnson, Laura D. (1982) *1974-1976 Air Quality by Location.* SEEDIS data document LBID-357-CD, January.

8. Johnson, Laura D., Merrill, D.W., and Selvin, S. (1982). *Predicting a continuous spatial variable from discrete point measurements.* LBL Report LBL-14235, March.

9. Johnson, Laura D. (1983) *The geographic and statistical analysis of air quality data in the United* States. LBL-Report LBL-16214, June.

10. Johnston, G.J. (1982) Probabilities of maximal deviations for nonparametric regression function estimates, *J. Mult. Analysis, 12, 402-414.*

11. Journel, A.G. and Huijbregts, C.J. (1978) *Mining Geostatistics.* Academic Press, London.

12. Merrill, Deane W. (1982) *Problems in Spatial Data Analysis.* LBL Report LBL-14047, February.

13. D.W. Merrill,Jr., Sacks, S.T., Selvin, S., Hollowell, C.D., and Winkelstein, Jr., W. (1978) *Populations-At-Risk-To-Air Pollution (PARAP): Data Base Description and Prototype Analysis.* LBL Report UCID-8039, August.

14. Merrill, D. (1981) *1974-1976 Air Quality for Individual Monitoring Stations.* SEEDIS data document LBID-357-AZ January.

15. Nadaraya, E. A. (1964). On estimating regression., *Theor. Probab. Appl. 9, 141-142.*

16. Brian D. Ripley (1981) *Spatial Statistics.* Wiley, October.

17. Scheffe', Henry (1959) *Analysis of Variance.* Wiley Publications in Statistics, pp. 57-58.

18. Selvin, S., Merrill, D., Kwok, L., and Sacks, S. (1981) *Ecologic Regression Analysis and the Study of the Influence of Air Quality on Mortality.* LBL Report LBL-12217, September.

19. Selvin, S., Sacks, S.T., Merrill, D.W., and Winkelstein, W. (1980) *The relationship between cancer incidence and two pollutants (total suspended particulate and carbon monoxide) for the San Francisco Bay Area.* LBL Report LBL-10847, June.

20. Stone, C. (1980) Optimal rates of convergence for nonparametric estimators. *Ann. Statist.. 8 '348-1360.*

21. Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society, Ser. B, 36, No. 2, 111-47.*

22. Stute, W. (1984) Asymptotic normality of nearest neighbor regression function estimates. *Ann. Statist., 12, 917-926.*

23. Wahba, G. and Wold, S. (1975) *A Completely Automatic French Curve: Fitting Spline Functions by Cross Validation.* Communications in Statistics, Vol. 4, no. 1, pp. 1-17.

24. Watson, G. S. (1964). Smooth regression analysis. *Sankhya. Ser. A Math. Sci. 26, 359-372.*

FIGURE 1. Weight as a function of distance from estimated point.

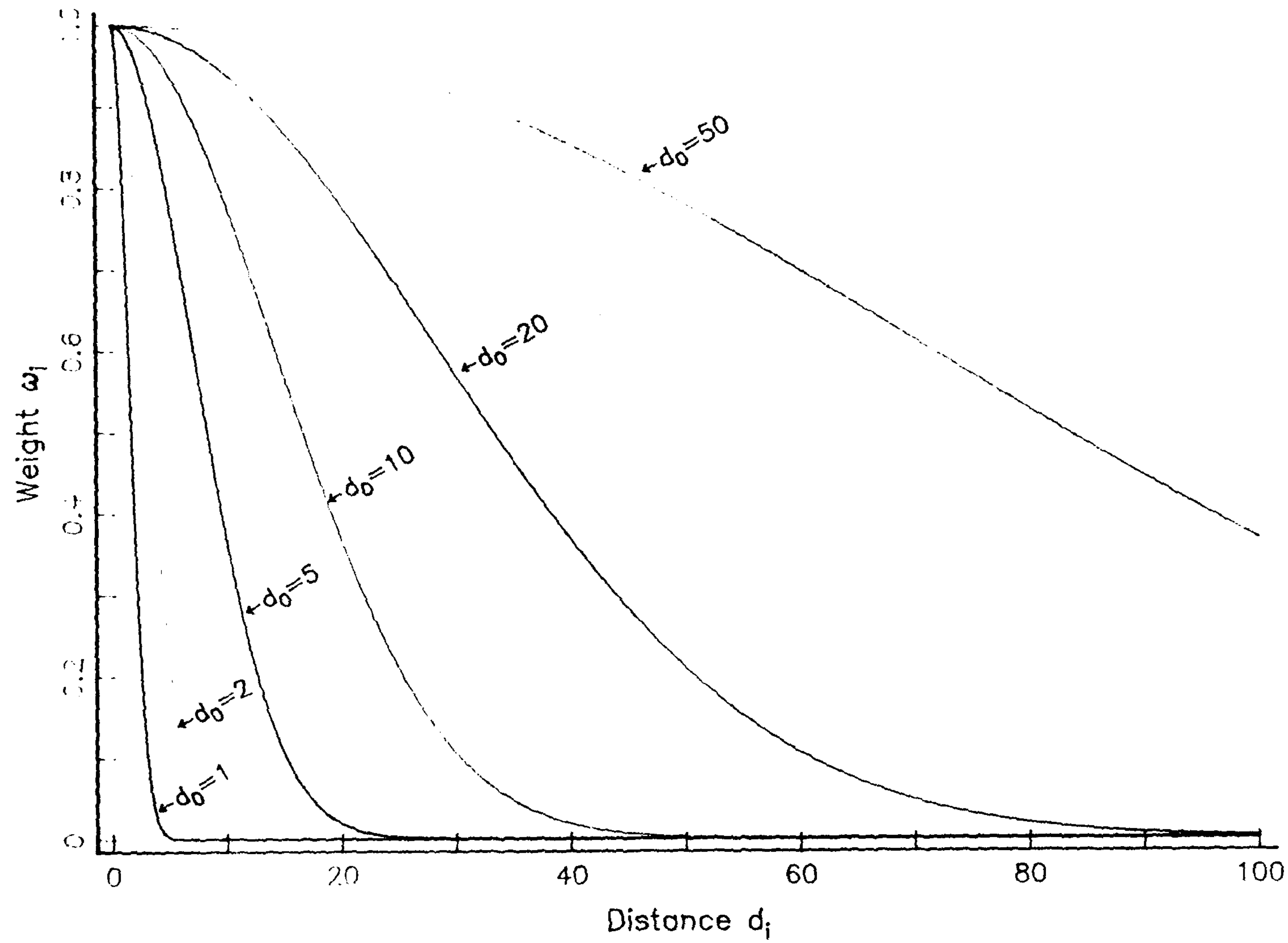


FIGURE 2. Mean Local Variance vs. Logarithm of Radius For Suspended Particulate in Ohio.

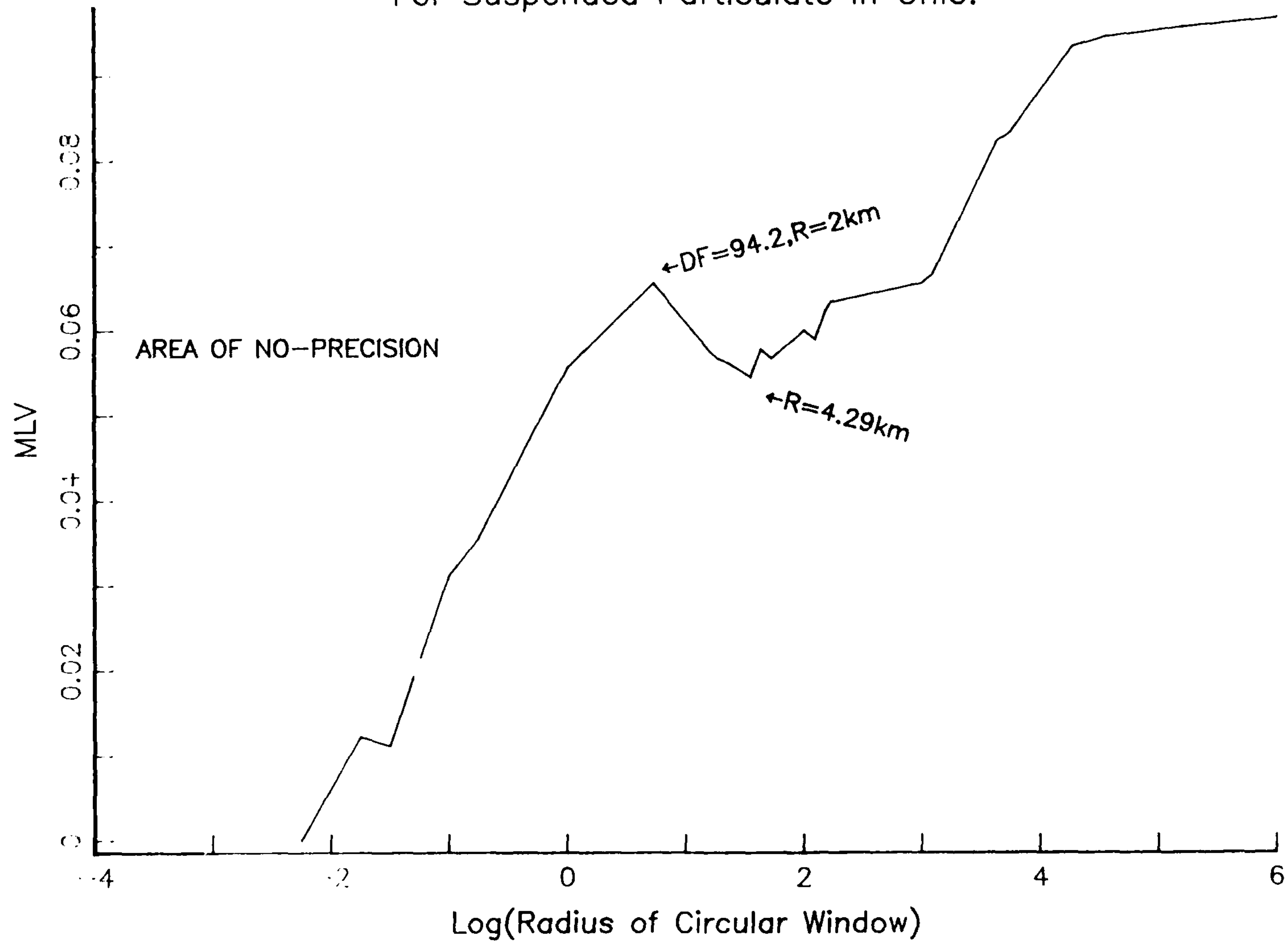


FIGURE 3. Mean Local Variance vs. Logarithm of Radius for Sulfur Dioxide in Ohio.

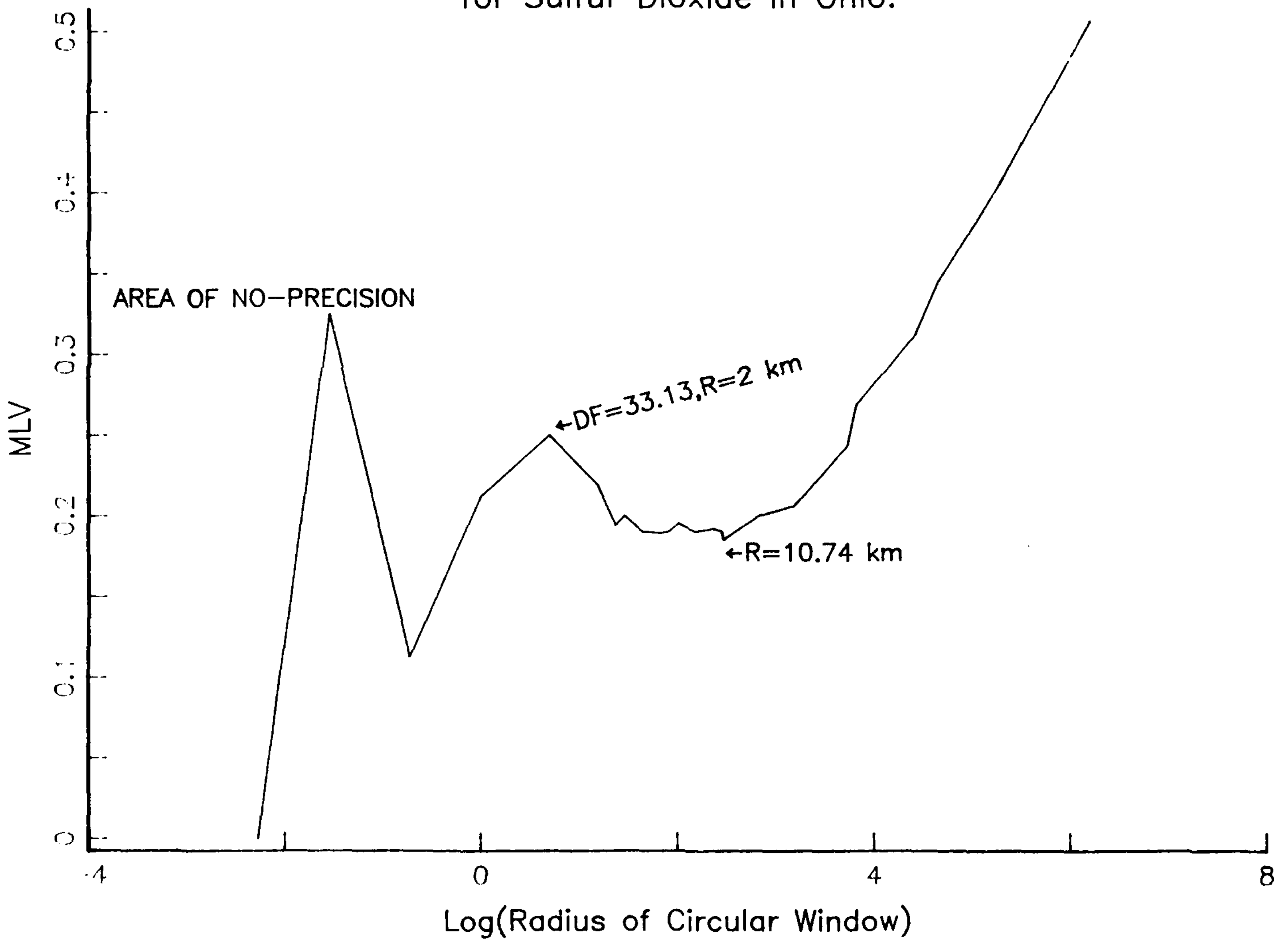


FIGURE 4. Mean Local Variance vs Log of Radius
for Suspended Particulate in NY state.

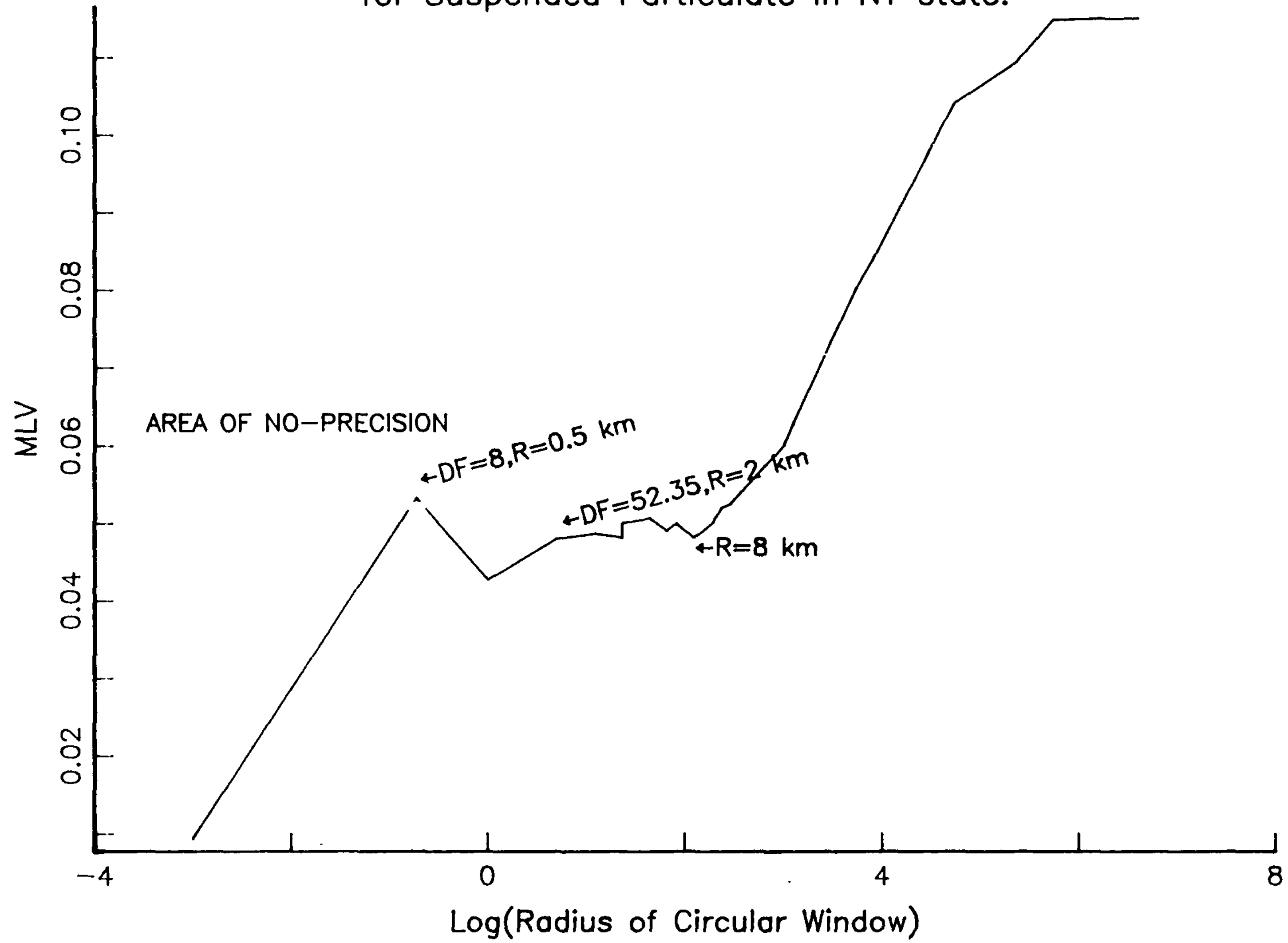


FIGURE 5. Mean Local Variance vs Log of Radius
For Sulfur Dioxide in NY State.

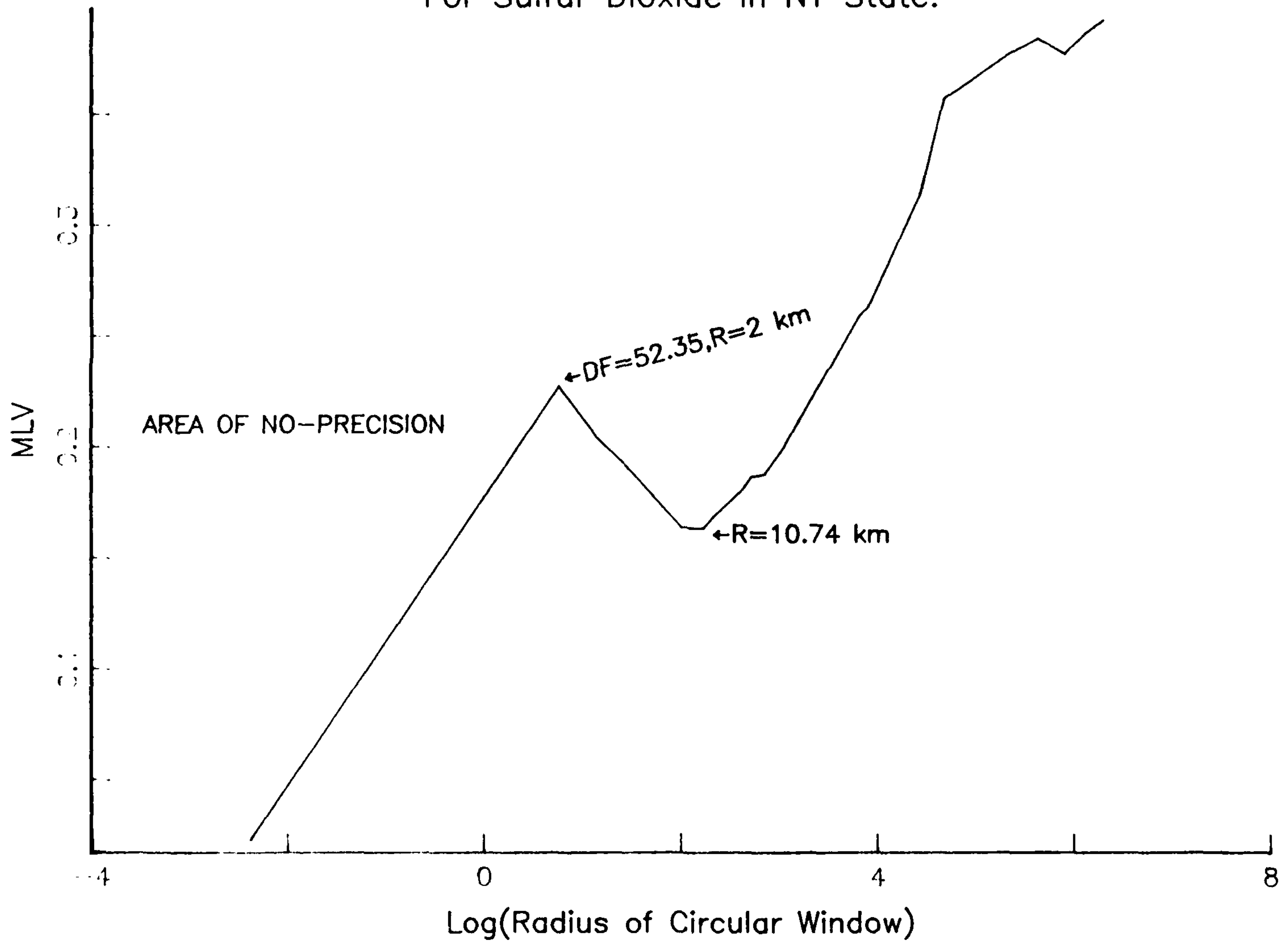


FIGURE 6. Mean Local Variance vs Log of Radius For Suspended Particulate in Florida State.

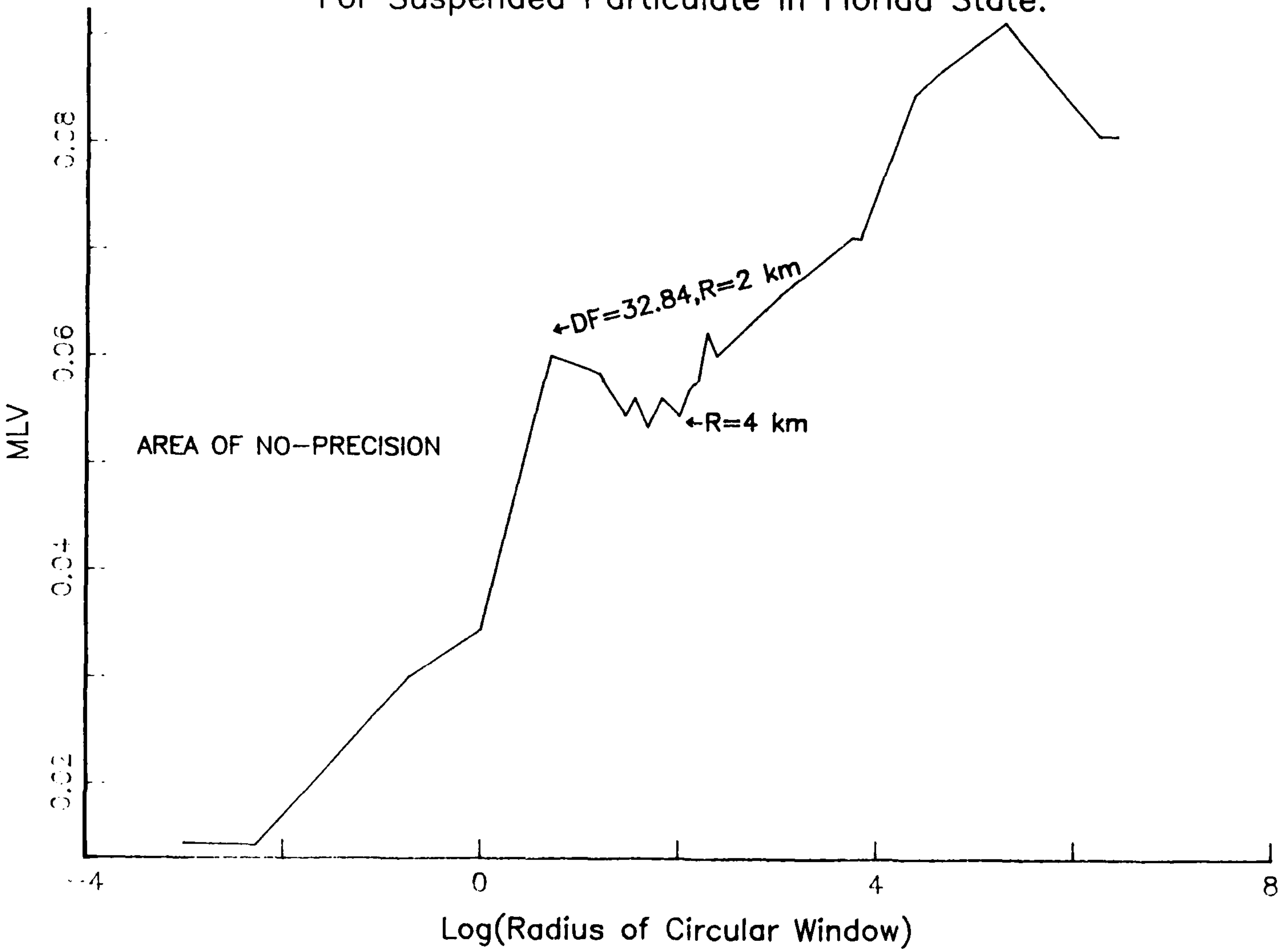


FIGURE 7. Mean Local Variance vs Log of Radius
For Sulfur Dioxide in Florida State.

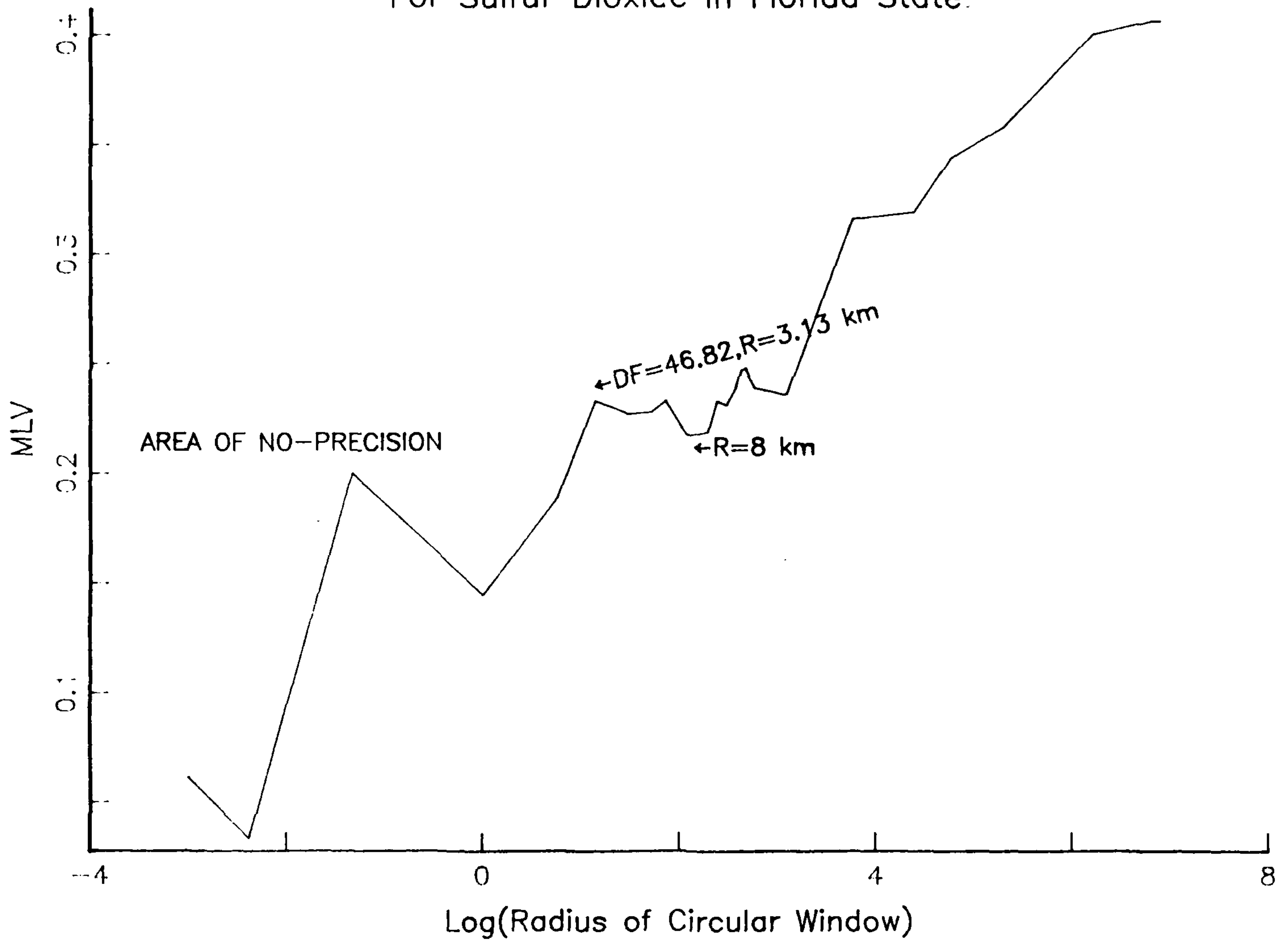
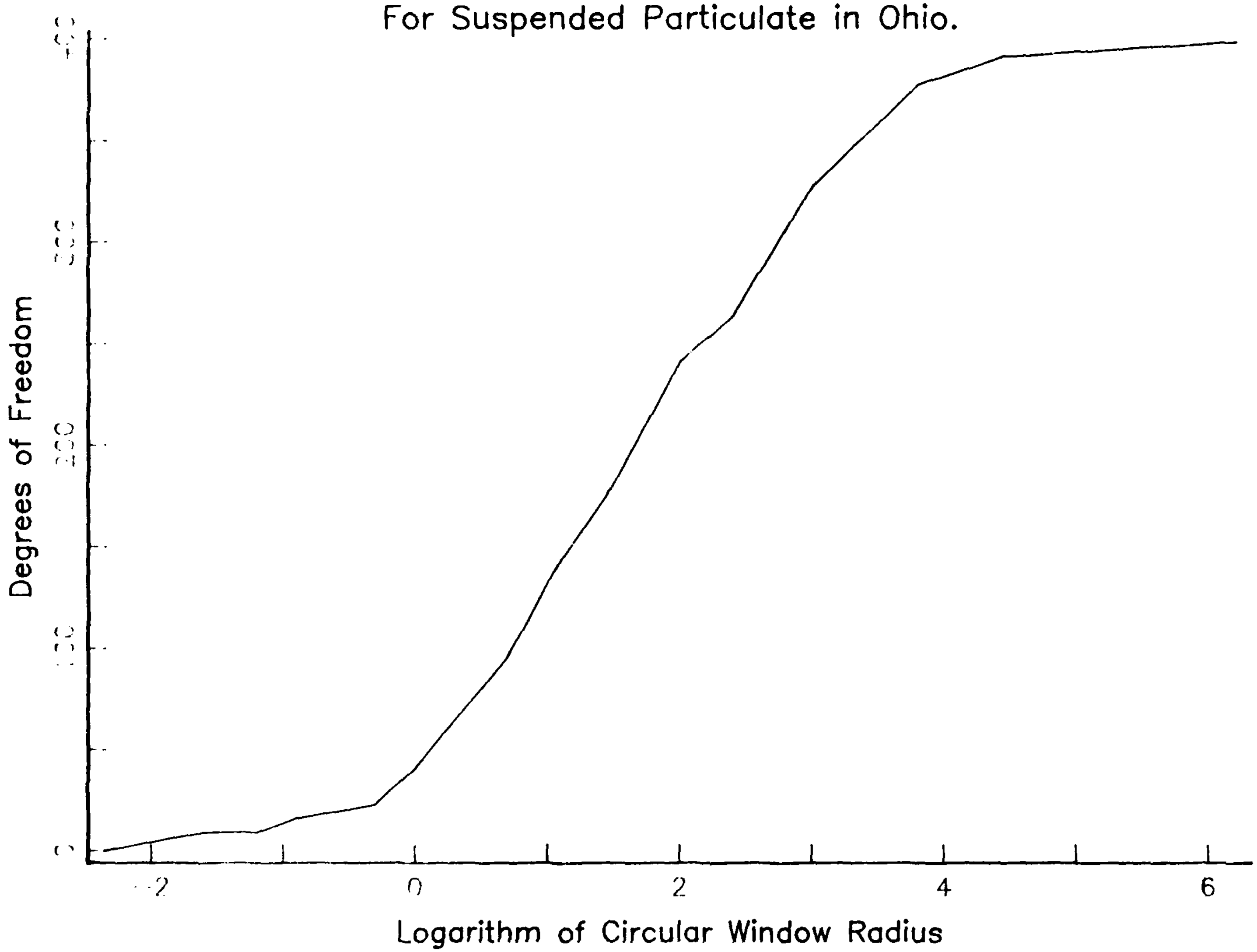


FIGURE 8. Degrees of Freedom vs Logarithm of Circle Radius
For Suspended Particulate in Ohio.



DISTRIBUTION LIST

	NO. OF COPIES
Library (Code 0142) Naval Postgraduate School Monterey, CA 93943-5000	2
Defense Technical Information Center Cameron Station Alexandria, VA 22314	2
Office of Research Administration (Code 012) Naval Postgraduate School Monterey, CA 93943-5000	1
Center for Naval Analyses 4401 Ford Avenue Alexandria, VA 22302-0268	1
Library (Code 55) Naval Postgraduate School Monterey, CA 93943-5000	1
Operations Research Center, Rm E40-164 Massachusetts Institute of Technology Attn: R. C. Larson and J. F. Shapiro Cambridge, MA 02139	1
Koh Peng Kong OA Branch, DSO Ministry of Defense Blk 29 Middlesex Road SINGAPORE 1024	1
Arthur P. Hurter, Jr. Professor and Chairman Dept of Industrial Engineering and Management Sciences Northwestern University Evanston, IL 60201-9990	1
Institute for Defense Analysis 1800 North Beauregard Alexandria, VA 22311	1

END

DATE

FILMED

8-88

DTIC